Reflected Text Analysis beyond Linguistics

DGfS-CL fall school

Nils Reiter,
nils.reiter@ims.uni-stuttgart.de

Sept. 9-13, 2019

Section 1

Summary

This course covered … I

Annotation

- Two purposes
  - Concept development
  - Creating reference data
- Annotation guidelines to mediate theory to annotators
  - Inter-Annotator Agreement is important
  - Other criteria as well

Evaluation

- Accuracy: Percentage of correctly classified instances
- Precision/recall: Per-class measures that reflect different kinds of errors
- Other metrics for other tasks

This course covered ... II

Automatisation

- Machine Learning: Algorithms to uncover patterns in data
  - Prediction model, training algorithm
- Classification: Categorising items (e.g., words $\rightarrow$ parts of speech)
- Decision trees
  - Well-established ML algorithm based on entropy
  - Highly transparent and well suited for categorial data
- Naive Bayes
  - Probabilistic model
  - Makes assumptions that clearly don't hold (in most cases)
- Shared tasks
  - Competitions to foster method/tool/guideline development

Part IV

What to do next

Using Machine Learning at Home

Section 2

Using Machine Learning at Home

Using Machine Learning at Home

The Task

What kind of problem do you want to solve?

- Classification: Items to classes
- Sequence labeling: Sequential items to classes
  - By taking previous decisions into account
  - Used in many NLP tasks!
- Regression: Predict numeric values
- Clustering: Data exploration

Using Machine Learning at Home

The Classes

What are the classes?

- Can humans distinguish between them clearly?
- Are there more training instances than classes?
- How specific are the classes to one document/data set?
  - Can we learn something generic from them?
- How are they distributed in the data/in the world?

Using Machine Learning at Home

The Data

- How large is the data set?
- Is it representative of the real world?
- Is it representative for the application?

Using Machine Learning at Home

The Features

Which features to use?

- Features need to be
  - Relevant for the target category
    - Your own judgement
    - Data analysis on a data sample: Association
  - Applicable to large portions of the instances
  - Extractable from the instances
    - How much time do you have?
    - How much preprocessing can you afford?
    - How reliable is the preprocessing?
- Extracting features: Main task for you
  - You'll have to write code

Processing Text

- Languages are different
  - German vs. English vs. Chinese

Processing Text

- Languages are different
  - German vs. English vs. Chinese
- Text types are different
  - Newspaper vs. blog vs. scientific articles

Processing Text

- Languages are different
  - German vs. English vs. Chinese
- Text types are different
  - Newspaper vs. blog vs. scientific articles
- Domains are different
  - Business vs. sports

Processing Text

- Languages are different
  - German vs. English vs. Chinese
- Text types are different
  - Newspaper vs. blog vs. scientific articles
- Domains are different
  - Business vs. sports

Processing Text

Differences are different

- Domain: Vocabulary
- Text types: Vocabulary, syntax, perspective, …
- Language: Syntax, vocabulary, semantics, sign systems, …

Processing Text

Ambiguity

Time flies like an arrow

Processing Text

Ambiguity

Time flies like an arrow

- Texts/sentences/words can be ambiguous
- How many different meanings does the sentence have?

Processing Text

Ambiguity

Angela saw the man with the binocular

Processing Text

Ambiguity

Angela saw the man with the binocular

- Ambiguity reflected in different syntactic readings
- PP attachment ambiguity
  - ‘see with the binocular’
  - ‘man with the binocular’

Processing Text

Processing text is hard

- NLP tools (e.g., Stanford Core NLP)
  - almost always supervised
  - trained on newspaper/Wikipedia/social media
- This may be what you need, but there's no guarantee.

Processing Text

Processing text is hard

- NLP tools (e.g., Stanford Core NLP)
  - almost always supervised
  - trained on newspaper/Wikipedia/social media
- This may be what you need, but there's no guarantee.
- Tools focus on linguistic layers (e.g., parts of speech or coreference)
  - Dependencies between layers exist!
  - PoS tagging errors lead to subsequent errors
    - This gap can be large Reiter (2014)

Processing Text

Processing text is hard

- NLP tools (e.g., Stanford Core NLP)
  - almost always supervised
  - trained on newspaper/Wikipedia/social media
- This may be what you need, but there's no guarantee.
- Tools focus on linguistic layers (e.g., parts of speech or coreference)
  - Dependencies between layers exist!
  - PoS tagging errors lead to subsequent errors
    - This gap can be large Reiter (2014)
- Technical text quality matters
  - 'Garbage in, garbage out'
  - OCR is not perfect

Section 3

Other Forms of Machine Learning for Text Analysis

Supervised vs. Unsupervised

Two strains of machine learning

Supervised Learning

- Goal: Replicate the gold standard
- Known classes
- Models trained on training data

$\rightarrow$ Classification

Supervised vs. Unsupervised

Two strains of machine learning

Supervised Learning

- Goal: Replicate the gold standard
- Known classes
- Models trained on training data

$\rightarrow$ Classification

Unsupervised Learning

- Goal: Identify groups of 'similar' items, similarity measured via features
  - Data exploration
- No gold standard, no training data

$\rightarrow$ Clustering

- Results not necessarily interpretable for humans!

Deep Learning / Neural Networks

- Relatively new development

Deep Learning / Neural Networks

- Relatively new development
- Major shift in workflow
  - No feature engineering
  - Input: Word embeddings and (very) large data sets
  - Benefits from very efficient linear algebra computation in graphics cards

Deep Learning / Neural Networks

- Relatively new development
- Major shift in workflow
  - No feature engineering
  - Input: Word embeddings and (very) large data sets
  - Benefits from very efficient linear algebra computation in graphics cards
- Downsides
  - Black box: Data in *n*-dimensional vector spaces is really hard to interpret
  - Not applicable below a certain amount of data
  - Ethical implications ($\rightarrow$ next week)
    - More severe than in 'classical machine learning'
    - 'Deep fakes', surveillance, ...

Section 4

Data & Annotation

Data

- Supervised ML needs (training/testing) data
- For text: Annotations!

Data

- Supervised ML needs (training/testing) data
- For text: Annotations!
- Corpus annotation
  - Tradition/established in computational linguistics
  - Explicitly marked linguistic categories
    - e.g., parts of speech (verb/noun/adjective/...)

Data

- Supervised ML needs (training/testing) data
- For text: Annotations!
- Corpus annotation
  - Tradition/established in computational linguistics
  - Explicitly marked linguistic categories
    - e.g., parts of speech (verb/noun/adjective/…)
- 'Distant supervision'
  - Generate training data from other sources (e.g., Wikipedia)
  - In many cases: More is better than higher quality
- Annotation as a by-product
  -

Getting Annotated Corpora

- LDC: Linguistic Data Consortium
  - https://www.ldc.upenn.edu
  - Intransparent business model …
- ELDA: European Language Resources Association
  - http://www.elra.info
- Open Access
  - Oxford Text Archive: http://ota.ox.ac.uk
  - Deutsches Textarchiv: http://www.deutschestextarchiv.de
  - TextGrid Repository: https://textgridrep.org
  - Project Gutenberg: http://www.gutenberg.org
  - Open Parallel cOrpUS: http://opus.nlpl.eu

Section 5

Resources

Continue Learning

- Coursera online course
  - Andrew Ng, Stanford University
  - https://www.coursera.org/learn/machine-learning
  - Lecture and exercises, generic (not only text/language)

Continue Learning

- Coursera online course
  - Andrew Ng, Stanford University
  - https://www.coursera.org/learn/machine-learning
  - Lecture and exercises, generic (not only text/language)
- Books
  - Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press, 1999
  - I. H. Witten and Eibe Frank. *Data Mining*. 2nd ed. Practical Machine Learning Tools and Techniques. Elsevier, Sept. 2005
  - Dan Jurafsky and James H. Martin. *Speech and Language Processing*. 2nd. Prentice Hall, 2008
  - Stefan Gries. *Quantitative Corpus Linguistics with R*. Routledge, 2009
  - Christoph Molnar. *Interpretable Machine Learning*. 2019. URL: `https://christophm.github.io/interpretable-ml-book/` (visited on 09/13/2019)

Summer schools and courses

- DGfS-CL:
  - Every two years, different institutes (in Germany)
  - Computational Linguistics
- ESU: European Summer University of Culture and Technology
  - Every summer in Leipzig, Germany
  - broad Digital Humanities
- DLLA: Deep Learning for Language Analysis
  - Cologne, Germany
  - Applied deep learning, written and spoken language

Start Coding

- You do not have to implement everything by yourself
  - Frameworks and APIs are faster, more tested, better documented
- Python
  - Natural Language Toolkit (NLTK): https://www.nltk.org
  - scikit-learn http://scikit-learn.org/
  - Industrial-Strength NLP https://spacy.io
- Java
  - Weka https://www.cs.waikato.ac.nz/ml/weka/
  - Mallet http://mallet.cs.umass.edu
  - Apache UIMA http://uima.apache.org
  - ClearTk http://cleartk.github.io/cleartk/
- R
  - caret https://topepo.github.io/caret/

Open Problems
(in my area of research)

- Narrative text analysis
  - Discourse structures beyond sentences
  - Content analysis, ‘plot’
  - Low data availability

Open Problems
(in my area of research)

- Narrative text analysis
  - Discourse structures beyond sentences
  - Content analysis, ‘plot’
  - Low data availability
- Small data analysis
  - No gigantic data sets in many CLS/DH areas
  - Different cost-benefit ratio

Open Problems
(in my area of research)

- Narrative text analysis
  - Discourse structures beyond sentences
  - Content analysis, ‘plot’
  - Low data availability
- Small data analysis
  - No gigantic data sets in many CLS/DH areas
  - Different cost-benefit ratio
- Semantic modelling for fictional texts/worlds
- Workflows for preventing misinterpretation of quantitative results
- …